

Atlantis Thinking Machines
Series Editor: K. - U. Kühnberger

Ben Goertzel
Cassio Pennachin
Nil Geisweiller

Engineering General Intelligence, Part 1

A Path to Advanced AGI via Embodied
Learning and Cognitive Synergy

Atlantis Thinking Machines

Volume 5

Series editor

Kai-Uwe Kühnberger, Osnabrück, Germany

For further volumes:
<http://www.atlantis-press.com>

Aims and Scope of the Series

This series publishes books resulting from theoretical research on and reproductions of general Artificial Intelligence (AI). The book series focuses on the establishment of new theories and paradigms in AI. At the same time, the series aims at exploring multiple scientific angles and methodologies, including results from research in cognitive science, neuroscience, theoretical and experimental AI, biology and from innovative interdisciplinary methodologies.

For more information on this series and our other book series, please visit our website at: www.atlantis-press.com/publications/books

AMSTERDAM—PARIS—BEIJING

ATLANTIS PRESS

Atlantis Press

29, avenue Laumière

75019 Paris, France

Ben Goertzel · Cassio Pennachin
Nil Geisweiller

Engineering General Intelligence, Part 1

A Path to Advanced AGI via Embodied
Learning and Cognitive Synergy

With contributions by the OpenCog Team



Ben Goertzel
G/F 51C Lung Mei Village
Tai Po
Hong Kong
People's Republic of China

Nil Geisweiller
Samokov
Bulgaria

Cassio Pennachin
Igenesis
Belo Horizonte, Minas Gerais
Brazil

ISSN 1877-3273

ISBN 978-94-6239-026-3

ISBN 978-94-6239-027-0 (eBook)

DOI 10.2991/978-94-6239-027-0

Library of Congress Control Number: 2013953280

© Atlantis Press and the authors 2014

This book, or any parts thereof, may not be reproduced for commercial purposes in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system known or to be invented, without prior permission from the Publisher.

Printed on acid-free paper

Preface

This is a large, two-part book with an even larger goal: To outline a practical approach to engineering software systems with general intelligence at the human level and ultimately beyond. Machines with flexible problem-solving ability, open-ended learning capability, creativity, and eventually their own kind of genius.

Part 1 of the book (Volume 5 in the Atlantis Thinking Machines book series), reviews various critical conceptual issues related to the nature of intelligence and mind. It then sketches the broad outlines of a novel, integrative architecture for Artificial General Intelligence (AGI) called CogPrime... and describes an approach for giving a young AGI system (CogPrime or otherwise) appropriate experience, so that it can develop its own smarts, creativity, and wisdom through its own experience. Along the way a formal theory of general intelligence is sketched, and a broad roadmap leading from here to human-level artificial intelligence. Hints are also given regarding how to eventually, potentially create machines advancing beyond human level—including some frankly futuristic speculations about strongly self-modifying AGI architectures with flexibility far exceeding that of the human brain.

Part 2 of the book (Volume 6 in the Atlantis Thinking Machines book series), then digs far deeper into the details of CogPrime's multiple structures, processes, and functions, culminating in a general argument as to why we believe CogPrime will be able to achieve general intelligence at the level of the smartest humans (and potentially greater), and a detailed discussion of how a CogPrime-powered virtual agent or robot would handle some simple practical tasks such as social play with blocks in a preschool context. It first describes the CogPrime software architecture and knowledge representation in detail; then reviews the cognitive cycle via which CogPrime perceives and acts in the world and reflects on itself; and next turns to various forms of learning: procedural, declarative (e.g., inference), simulative, and integrative. Methods of enabling natural language functionality in CogPrime are then discussed; and then the volume concludes with a chapter summarizing the argument that CogPrime can lead to human-level (and eventually perhaps greater) AGI, and a chapter giving a thought experiment describing the internal dynamics via which a completed CogPrime system might solve the problem of obeying the request "Build me something with blocks that I haven't seen before."

The chapters here are written to be read in linear order—and if consumed thus, they tell a coherent story about how to get from here to advanced AGI.

However, we suggest the impatient reader may wish to take a quick look at the final chapter of Part 2, after reading [Chaps. 1–3](#) of Part 1. This final chapter gives a broad overview of why we think the CogPrime design will work, in a way that depends on the technical details of the previous chapters, but (we believe) not so sensitively as to be incomprehensible without them.

This is admittedly an unusual sort of book, mixing demonstrated conclusions with unproved conjectures in a complex way, all oriented toward an extraordinarily ambitious goal. Further, the chapters are somewhat variant in their levels of detail—some very nitty-gritty, some more high level, with much of the variation due to how much concrete work has been done on the topic of the chapter at time of writing. However, it is important to understand that the ideas presented here are not mere armchair speculation—they are currently being used as the basis for an open-source software project called OpenCog, which is being worked on by software developers around the world. Right now OpenCog embodies only a percentage of the overall CogPrime design as described here. But if OpenCog continues to attract sufficient funding or volunteer interest, then the ideas presented in these volumes will be validated or refuted via practice (As a related note: here and there in this book, we will refer to the “current” CogPrime implementation (in the OpenCog framework); in all cases this refers to OpenCog as of late 2013).

To state one believes one knows a workable path to creating a human-level (and potentially greater) general intelligence is to make a dramatic statement, given the conventional way of thinking about the topic in the contemporary scientific community. However, we feel that once a little more time has passed, the topic will lose its drama (if not its interest and importance), and it will be widely accepted that there are *many* ways to create intelligent machines—some simpler and some more complicated; some more brain-like or human-like and some less so; some more efficient and some more wasteful of resources; etc. We have little doubt that, from the perspective of AGI science 50 or 100 years hence (and probably even 10–20 years hence), the specific designs presented here will seem awkward, messy, inefficient, and circuitous in various respects. But that is how science and engineering progress. Given the current state of knowledge and understanding, having any concrete, comprehensive design, and plan for creating AGI is a significant step forward; and it is in this spirit that we present here our thinking about the CogPrime architecture and the nature of general intelligence.

In the words of Sir Edmund Hillary, the first to scale Everest: “Nothing Venture, Nothing Win.”

Prehistory of the Book

The writing of this book began in earnest in 2001, at which point it was informally referred to as “The Novamente Book.” The original “Novamente Book” manuscript ultimately got too big for its own britches, and subdivided into a number of different works—*The Hidden Pattern* (Goertzel 2006), a philosophy of mind book

published in 2006; *Probabilistic Logic Networks* (Goertzel et al. 2008), a more technical work published in 2008; *Real World Reasoning* (Goertzel et al. 2011), a sequel to *Probabilistic Logic Networks* published in 2011; and the two parts of this book.

The ideas described in this book have been the collaborative creation of multiple overlapping communities of people over a long period of time. The vast bulk of the writing here was done by Ben Goertzel; but Cassio Pennachin and Nil Geisweiller made sufficient writing, thinking, and editing contributions over the years to more than merit their inclusion as coauthors. Further, many of the chapters here have coauthors beyond the three main coauthors of the book; and the set of chapter coauthors does not exhaust the set of significant contributors to the ideas presented.

The core concepts of the CogPrime design and the underlying theory were conceived by Ben Goertzel in the period 1995–1996 when he was a Research Fellow at the University of Western Australia; but those early ideas have been elaborated and improved by many more people than can be listed here (as well as by Ben’s ongoing thinking and research). The collaborative design process ultimately resulting in CogPrime started in 1997 when Intelligenesis Corp. was formed—the Webmind AI Engine created in Intelligenesis’s research group during 1997–2001 was the predecessor to the Novamente Cognition Engine created at Novamente LLC during 2001–2008, which was the predecessor to CogPrime.

Online Appendices

Just one more thing before getting started! This book originally had even more chapters than the ones currently presented in Parts 1 and 2. In order to decrease length and increase focus, however, a number of chapters dealing with peripheral—yet still relevant and interesting—matters were moved to online appendices. These may be downloaded in a single PDF file at http://goertzel.org/engineering_general_intelligence_appendices_B-H.pdf. The titles of these appendices are:

- Appendix A: Possible Worlds Semantics and Experiential Semantics
- Appendix B: Steps Toward a Formal Theory of Cognitive Structure and Dynamics
- Appendix C: Emergent Reflexive Mental Structures
- Appendix D: GOLEM: Toward an AGI Meta-Architecture Enabling Both Goal Preservation and Radical Self-Improvement
- Appendix E: Lojban++: A Novel Linguistic Mechanism for Teaching AGI Systems
- Appendix F: Possible Worlds Semantics and Experiential Semantics
- Appendix G: PLN and the Brain
- Appendix H: Propositions About Environments in Which CogPrime Components Are Useful

None of these are critical to understanding the key ideas in the book, which is why they were relegated to online appendices. However, reading them will deepen your understanding of the conceptual and formal perspectives underlying the CogPrime design. These appendices are referred to here and there in the text of the main book.

References

- B. Goertzel, *The Hidden Pattern* (Brown Walker, Boca Raton, 2006)
- B. Goertzel, M. Ikle, I. Goertzel, A. Heljakka, *Probabilistic Logic Networks* (Springer, Heidelberg, 2008)
- B. Goertzel, N. Geisweiller, L. Coelho, P. Janicic, C. Pennachin, *Real World Reasoning* (Atlantis, Oxford, 2011)

Acknowledgments

For sake of simplicity, this acknowledgments section is presented from the perspective of the primary author, Ben Goertzel. Ben will thus begin by expressing his thanks to his primary coauthors, Cassio Pennachin (Collaborator since 1998) and Nil Geisweiller (Collaborator since 2005). Without outstandingly insightful, deep-thinking colleagues like you, the ideas presented here—let alone the book itself—would not have developed nearly as effectively as what has happened. Similar thanks also go to the other OpenCog Collaborators who have coauthored various chapters of the book.

Beyond the coauthors, huge gratitude must also be extended to everyone who has been involved with the OpenCog project, and/or was involved in Novamente LLC and Webmind Inc. before that. We are grateful to all of you for your collaboration and intellectual companionship!

Building a thinking machine is a huge project, too big for any one human; it will take a team and I'm happy to be part of a great one. It is through the genius of human collectives, going beyond any individual human mind, that genius machines are going to be created.

A tiny, incomplete sample from the long list of those others deserving thanks is:

- Ken Silverman and Gwendalin Qi Aranya (formerly Gwen Goertzel), both of whom listened to me talk at inordinate length about many of the ideas presented here a long, long time before anyone else was interested in listening. Ken and I schemed some AGI designs at Simon's Rock College in 1983, years before we worked together on the Webmind AI Engine.
- Allan Combs, who got me thinking about consciousness in various different ways, at a very early point in my career. I'm very pleased to still count Allan as a friend and sometime Collaborator! Fred Abraham as well, for introducing me to the intersection of chaos theory and cognition, with a wonderful flair. George Christos, a deep AI/math/physics thinker from Perth, for reawakening my interest in attractor neural nets and their cognitive implications, in the mid-1990s.
- All of the 130 staff of Webmind Inc. during 1998–2001 while that remarkable, ambitious, peculiar AGI-oriented firm existed. Special shout-outs to the “Voice of Reason” Pei Wang and the “Siberian Madmind” Anton Kolonin, Mike Ross, Cate Hartley, Karin Verspoor, and the tragically prematurely deceased Jeff

Pressing (compared to whom we are all mental midgets), who all made serious conceptual contributions to my thinking about AGI. Lisa Pazer and Andy Siciliano who made Webmind happen on the business side. And of course Cassio Pennachin, a coauthor of this book; and Ken Silverman, who coarchitected the whole Webmind system and vision with me from the start.

- The Webmind Diehards, who helped begin the Novamente project that succeeded Webmind beginning in 2001: Cassio Pennachin, Stephan Vladimir Bugaj, Takuo Henmi, Matthew Ikle', Thiago Maia, Andre Senna, Guilherme Lamacie, and Saulo Pinto.
- Those who helped get the Novamente project off the ground and keep it progressing over the years, including some of the Webmind Diehards and also Moshe Looks, Bruce Klein, Izabela Lyon Freire, Chris Poulin, Murilo Queiroz, Predrag Janicic, David Hart, Ari Heljakka, Hugo Pinto, Deborah Duong, Paul Prueitt, Glenn Tarbox, Nil Geisweiller and Cassio Pennachin (the coauthors of this book), Sibley Verbeck, Jeff Reed, Pejman Makhfi, Welter Silva, Lukasz Kaiser, and more.
- Izabela Lyon Freire, for endless wide-ranging discussions on AI, philosophy of mind and related topics, which helped concretize many of the ideas here.
- David Hart, for his various, inspired IT and editing help over the years, including lots of editing and formatting help on an earlier version of this book which took the form of a Wikibook.
- A group of volunteers who helped proofread a draft of this manuscript, including Brett Woodward, Terry Ward, Dario Garcia Gasulla, Tom Heiman as well as others. Tim Josling's efforts in this regard were especially thorough and helpful.
- All those who have helped with the OpenCog system, including Linas Vepstas, Joel Pitt, Jared Wigmore/Jade O'Neill, Zhenhua Cai, Deheng Huang, Shujing Ke, Lake Watkins, Alex van der Peet, Samir Araujo, Fabricio Silva, Yang Ye, Shuo Chen, Michel Drenthe, Ted Sanders, Gustavo Gama and of course Nil, and Cassio again. Tyler Emerson and Eliezer Yudkowsky, for choosing to have the Singularity Institute for AI (now MIRI) provide seed funding for OpenCog.
- The numerous members of the AGI community who have tossed around AGI ideas with me since the first AGI conference in 2006, including but definitely not limited to: Stan Franklin, Juergen Schmidhuber, Marcus Hutter, Kai-Uwe Kühnberger, Stephen Reed, Blerim Enruli, Kristinn Thorisson, Joscha Bach, Abram Demski, Itamar Arel, Mark Waser, Randal Koene, Paul Rosenbloom, Zhongzhi Shi, Steve Omohundro, Bill Hibbard, Eray Ozkural, Brandon Rohrer, Ben Johnston, John Laird, Shane Legg, Selmer Bringsjord, Anders Sandberg, Alexei Samsonovich, Wlodek Duch, and more.
- The inimitable "Artilect Warrior" Hugo de Garis, who (when he was working at Xiamen University) got me started working on AGI in the Orient (and introduced me to my wife Ruiting in the process). And Changle Zhou, who brought Hugo to Xiamen and generously shared his brilliant research students with Hugo and me. And Min Jiang, Collaborator of Hugo and Changle, a deep AGI thinker who is helping with OpenCog theory and practice at time of writing.

- Gino Yu, who got me started working on AGI here in Hong Kong, where I am living at time of writing. As of 2013 the bulk of OpenCog work is occurring in Hong Kong via a research grant that Gino and I obtained together.
- Dan Stoicescu, whose funding helped Novamente through some tough times.
- Jeffrey Epstein, whose visionary funding of my AGI research has helped me through a number of tight spots over the years. At time of writing, Jeffrey is helping support the OpenCog Hong Kong project.
- Zeger Karssen, Founder of Atlantis Press, who conceived the *Thinking Machines* book series in which this book appears, and who has been a strong supporter of the AGI conference series from the beginning.
- My wonderful wife Ruiting Lian, a source of fantastic amounts of positive energy for me since we became involved 4 years ago. Ruiting has listened to me discuss the ideas contained here time and time again, often with judicious and insightful feedback (as she is an excellent AI Researcher in her own right); and has been wonderfully tolerant of me diverting numerous evenings and weekends to getting this book finished (as well as to other AGI-related pursuits). And my parents Ted and Carol and kids Zar, Zeb, and Zade, who have also indulged me in discussions on many of the themes discussed here on countless occasions! And my dear, departed grandfather Leo Zwell, for getting me started in science.
- Crunchkin and Pumpkin, for regularly getting me away from the desk to stroll around the village where we live; many of my best ideas about AGI and other topics have emerged while walking with my furry four-legged family members.

September 2013

Ben Goertzel

Contents

1	Introduction	1
1.1	AI Returns to Its Roots	1
1.2	AGI Versus Narrow AI	2
1.3	CogPrime.	3
1.4	The Secret Sauce	4
1.5	Extraordinary Proof?	5
1.6	Potential Approaches to AGI	6
1.6.1	Build AGI from Narrow AI	6
1.6.2	Enhancing Chatbots.	7
1.6.3	Emulating the Brain	7
1.6.4	Evolve an AGI	8
1.6.5	Derive an AGI Design Mathematically	8
1.6.6	Use Heuristic Computer Science Methods	9
1.6.7	Integrative Cognitive Architecture.	9
1.6.8	Can Digital Computers Really Be Intelligent?	10
1.7	Five Key Words	11
1.7.1	Memory and Cognition in CogPrime.	11
1.8	Virtually and Robotically Embodied AI.	13
1.9	Language Learning	13
1.10	AGI Ethics.	14
1.11	Structure of the Book	15
1.12	Key Claims of the Book	15

Part I Overview of the CogPrime Architecture

2	A Brief Overview of CogPrime	21
2.1	Introduction	21
2.2	High-Level Architecture of CogPrime	21
2.3	Current and Prior Applications of OpenCog.	23
2.3.1	Transitioning from Virtual Agents to a Physical Robot	25

2.4	Memory Types and Associated Cognitive Processes in CogPrime.	30
2.4.1	Cognitive Synergy in PLN.	32
2.5	Goal-Oriented Dynamics in CogPrime.	35
2.6	Analysis and Synthesis Processes in CogPrime.	36
2.7	Conclusion.	40
3	Build Me Something I Haven't Seen: A CogPrime Thought Experiment.	41
3.1	Introduction	41
3.2	Roles of Selected Cognitive Processes.	42
3.3	A Semi-Narrative Treatment	54
3.4	Conclusion.	57
 Part II Artificial and Natural General Intelligence		
4	What is Human-Like General Intelligence?	61
4.1	Introduction	61
4.1.1	What is General Intelligence?	62
4.1.2	What is Human-Like General Intelligence?	62
4.2	Commonly Recognized Aspects of Human-Like Intelligence	63
4.3	Further Characterizations of Human-Like Intelligence.	66
4.3.1	Competencies Characterizing Human-Like Intelligence	66
4.3.2	Gardner's Theory of Multiple Intelligences	67
4.3.3	Newell's Criteria for a Human Cognitive Architecture	68
4.3.4	Intelligence and Creativity	69
4.4	Preschool as a View into Human-Like General Intelligence	70
4.4.1	Design for an AGI Preschool	72
4.5	Integrative and Synergetic Approaches to Artificial General Intelligence	74
4.5.1	Achieving Human-Like Intelligence via Cognitive Synergy	75
5	A Patternist Philosophy of Mind	77
5.1	Introduction	77
5.2	Some Patternist Principles	78
5.3	Cognitive Synergy	83

5.4	The General Structure of Cognitive Dynamics: Analysis and Synthesis	85
5.4.1	Component-Systems and Self-Generating Systems	85
5.4.2	Analysis and Synthesis	87
5.4.3	The Dynamic of Iterative Analysis and Synthesis	90
5.4.4	Self and Focused Attention as Approximate Attractors of the Dynamic of Iterated Forward-Analysis	91
5.4.5	Conclusion	94
5.5	Perspectives on Machine Consciousness	95
5.6	Postscript: Formalizing Pattern	96
6	Brief Survey of Cognitive Architectures	101
6.1	Introduction	101
6.2	Symbolic Cognitive Architectures	103
6.2.1	SOAR	104
6.2.2	ACT-R	105
6.2.3	Cyc and Texai	107
6.2.4	NARS	107
6.2.5	GLAIR and SNePS	108
6.3	Emergentist Cognitive Architectures	109
6.3.1	DeSTIN: A Deep Reinforcement Learning Approach to AGI	111
6.3.2	Developmental Robotics Architectures	117
6.4	Hybrid Cognitive Architectures	119
6.4.1	Neural Versus Symbolic; Global Versus Local	121
6.5	Globalist Versus Localist Representations	124
6.5.1	CLARION	125
6.5.2	The Society of Mind and the Emotion Machine	126
6.5.3	DUAL	126
6.5.4	4D/RCS	128
6.5.5	PolyScheme	130
6.5.6	Joshua Blue	130
6.5.7	LIDA	131
6.5.8	The Global Workspace	131
6.5.9	The LIDA Cognitive Cycle	132
6.5.10	Psi and MicroPsi	136
6.5.11	The Emergence of Emotion in the Psi Model	139
6.5.12	Knowledge Representation, Action Selection and Planning in Psi	140
6.5.13	Psi Versus CogPrime	142

7	A Generic Architecture of Human-Like Cognition	143
7.1	Introduction	143
7.2	Key Ingredients of the Integrative Human-Like Cognitive Architecture Diagram	144
7.3	An Architecture Diagram for Human-Like General Intelligence	146
7.4	Interpretation and Application of the Integrative Diagram . .	152

Part III Toward a General Theory of General Intelligence

8	A Formal Model of Intelligent Agents	157
8.1	Introduction	157
8.2	A Simple Formal Agents Model (SRAM)	158
8.2.1	Goals	159
8.2.2	Memory Stores	160
8.2.3	The Cognitive Schematic	162
8.3	Toward a Formal Characterization of Real-World General Intelligence	163
8.3.1	Biased Universal Intelligence	165
8.3.2	Connecting Legg and Hutter’s Model of Intelligent Agents to the Real World	166
8.3.3	Pragmatic General Intelligence	167
8.3.4	Incorporating Computational Cost	168
8.3.5	Assessing the Intelligence of Real-World Agents. . .	169
8.4	Intellectual Breadth: Quantifying the Generality of an Agent’s Intelligence	171
8.5	Conclusion	172
9	Cognitive Synergy	173
9.1	Introduction	173
9.2	Cognitive Synergy	174
9.3	Cognitive Synergy in CogPrime	177
9.3.1	Cognitive Processes in CogPrime	177
9.4	Some Critical Synergies	182
9.5	Cognitive Synergy for Procedural and Declarative Learning.	182
9.5.1	Cognitive Synergy in MOSES	187
9.5.2	Cognitive Synergy in PLN	190
9.6	Is Cognitive Synergy Tricky?	191
9.6.1	The Puzzle: Why Is It So Hard to Measure Partial Progress Toward Human-Level AGI?	192
9.6.2	A Possible Answer: Cognitive Synergy Is Tricky! . .	193
9.6.3	Conclusion	194

10	General Intelligence in the Everyday Human World	197
10.1	Introduction	197
10.2	Some Broad Properties of the Everyday World that Help Structure Intelligence	198
10.3	Embodied Communication	199
10.3.1	Generalizing the Embodied Communication Prior.	202
10.4	Naive Physics.	203
10.4.1	Objects, Natural Units and Natural Kinds	204
10.4.2	Events, Processes and Causality	204
10.4.3	Stuffs, States of Matter, Qualities	205
10.4.4	Surfaces, Limits, Boundaries, Media	205
10.4.5	What Kind of Physics is Needed to Foster Human-Like Intelligence?	206
10.5	Folk Psychology	208
10.5.1	Motivation, Requiredness, Value.	208
10.6	Body and Mind	208
10.6.1	The Human Sensorium	209
10.6.2	The Human Body's Multiple Intelligences	210
10.7	The Extended Mind and Body	213
10.8	Conclusion.	213
11	A Mind-World Correspondence Principle	215
11.1	Introduction	215
11.2	What Might a General Theory of General Intelligence Look Like?	216
11.3	Steps Toward A (Formal) General Theory of General Intelligence	217
11.4	The Mind-World Correspondence Principle	219
11.5	How Might the Mind-World Correspondence Principle Be Useful?	220
11.6	Conclusion.	221

Part IV Cognitive and Ethical Development

12	Stages of Cognitive Development.	225
12.1	Introduction	225
12.2	Piagetan Stages in the Context of a General Systems Theory of Development.	226
12.3	Piaget's Theory of Cognitive Development	227
12.3.1	Perry's Stages.	231
12.3.2	Keeping Continuity in Mind.	232

12.4	Piaget's Stages in the Context of Uncertain Inference	232
12.4.1	The Infantile Stage	234
12.4.2	The Concrete Stage	236
12.4.3	The Formal Stage	240
12.4.4	The Reflexive Stage	242
13	The Engineering and Development of Ethics	245
13.1	Introduction	245
13.2	Review of Current Thinking on the Risks of AGI.	246
13.3	The Value of an Explicit Goal System	249
13.4	Ethical Synergy	251
13.4.1	Stages of Development of Declarative Ethics	252
13.4.2	Stages of Development of Empathic Ethics	255
13.4.3	An Integrative Approach to Ethical Development	257
13.4.4	Integrative Ethics and Integrative AGI.	259
13.5	Clarifying the Ethics of Justice: Extending the Golden Rule in to a Multifactorial Ethical Model	261
13.5.1	The Golden Rule and the Stages of Ethical Development	264
13.5.2	The Need for Context-Sensitivity and Adaptiveness in Deploying Ethical Principles.	266
13.6	The Ethical Treatment of AGIs	269
13.6.1	Possible Consequences of Depriving AGIs of Freedom.	271
13.6.2	AGI Ethics as Boundaries Between Humans and AGIs Become Blurred	272
13.7	Possible Benefits of Closely Linking AGIs to the Global Brain	273
13.7.1	The Importance of Fostering Deep, Consensus-Building Interactions Between People with Divergent Views	275
13.8	Possible Benefits of Creating Societies of AGIs	277
13.9	AGI Ethics as Related to Various Future Scenarios.	278
13.9.1	Capped Intelligence Scenarios	278
13.9.2	Superintelligent AI: Soft-Takeoff Scenarios	279
13.9.3	Superintelligent AI: Hard-Takeoff Scenarios	280
13.9.4	Global Brain Mindplex Scenarios	281
13.10	Conclusion: Eight Ways to Bias AGI Toward Friendliness	284
13.10.1	Encourage Measured Co-Advancement of AGI Software and AGI Ethics Theory	286
13.10.2	Develop Advanced AGI Sooner Not Later	286

Part V Networks for Explicit and Implicit Knowledge Representation

14	Local, Global and Glocal Knowledge Representation	291
14.1	Introduction	291
14.2	Localized Knowledge Representation Using Weighted, Labeled Hypergraphs.	292
14.2.1	Weighted, Labeled Hypergraphs	292
14.3	Atoms: Their Types and Weights	293
14.3.1	Some Basic Atom Types	294
14.3.2	Variable Atoms.	296
14.3.3	Logical Links	297
14.3.4	Temporal Links	299
14.3.5	Associative Links	300
14.3.6	Procedure Nodes	301
14.3.7	Links for Special External Data Types	302
14.3.8	Truth Values and Attention Values	302
14.4	Knowledge Representation via Attractor Neural Networks. . .	303
14.4.1	The Hopfield Neural Net Model	303
14.4.2	Knowledge Representation via Cell Assemblies . .	304
14.5	Neural Foundations of Learning	305
14.5.1	Hebbian Learning	305
14.5.2	Virtual Synapses and Hebbian Learning Between Assemblies	306
14.5.3	Neural Darwinism.	307
14.6	Glocal Memory	308
14.6.1	A Semi-Formal Model of Glocal Memory	310
14.6.2	Glocal Memory in the Brain.	312
14.6.3	Glocal Hopfield Networks	315
14.6.4	Neural-Symbolic Glocality in CogPrime	317
15	Representing Implicit Knowledge via Hypergraphs	319
15.1	Introduction	319
15.2	Key Vertex and Edge Types	320
15.3	Derived Hypergraphs.	320
15.3.1	SMEPH Vertices.	321
15.3.2	SMEPH Edges	322
15.4	Implications of Patternist Philosophy for Derived Hypergraphs of Intelligent Systems.	323
15.4.1	SMEPH Principles in CogPrime	324

16	Emergent Networks of Intelligence	327
16.1	Introduction	327
16.2	Small World Networks	328
16.3	Dual Network Structure	329
16.3.1	Hierarchical Networks	329
16.3.2	Associative, Heterarchical Networks	332
16.3.3	Dual Networks	333
Part VI A Path to Human-Level AGI		
17	AGI Preschool	337
17.1	Introduction	337
17.1.1	Contrast to Standard AI Evaluation Methodologies	338
17.2	Elements of Preschool Design	340
17.3	Elements of Preschool Curriculum	341
17.3.1	Preschool in the Light of Intelligence Theory	341
17.4	Task-Based Assessment in AGI Preschool	343
17.5	Beyond Preschool	346
17.6	Issues with Virtual Preschool Engineering	347
17.6.1	Integrating Virtual Worlds with Robot Simulators	349
17.6.2	BlocksNBeads World	349
18	A Preschool-Based Roadmap to Advanced AGI	355
18.1	Introduction	355
18.2	Measuring Incremental Progress Toward Human-Level AGI	356
18.3	Conclusion	364
19	Advanced Self-Modification: A Possible Path to Superhuman AGI	365
19.1	Introduction	365
19.2	Cognitive Schema Learning	367
19.3	Self-Modification via Supercompilation	368
19.3.1	Three Aspects of Supercompilation	370
19.3.2	Supercompilation for Goal-Directed Program Modification	371
19.4	Self-Modification via Theorem-Proving	372
Appendix A: Glossary		375
 Index		405

Acronyms

AA	Attention Allocation
ADF	Automatically Defined Function (in the context of Genetic Programming)
AF	Attentional Focus
AGI	Artificial General Intelligence
AV	Attention Value
BD	Behavior Description
C-space	Configuration Space
CBV	Coherent Blended Volition
CEV	Coherent Extrapolated Volition
CGGP	Contextually Guided Greedy Parsing
CSDLN	Compositional Spatiotemporal Deep Learning Network
CT	Combo Tree
ECAN	Economic Attention Network
ECP	Embodied Communication Prior
EPW	Experiential Possible Worlds (semantics)
FCA	Formal Concept Analysis
FI	Fisher Information
FIM	Frequent Itemset Mining
FOI	First Order Inference
FOPL	First Order Predicate Logic
FOPLN	First Order PLN
FS-MOSES	Feature Selection MOSES (i.e. MOSES with feature selection integrated a la LIFES)
GA	Genetic Algorithms
GB	Global Brain
GEOP	Goal Evaluator Operating Procedure (in a GOLEM context)
GIS	Geospatial Information System
GOLEM	Goal-Oriented LEarning Meta-architecture
GP	Genetic Programming
HOI	Higher-Order Inference
HOPLN	Higher-Order PLN

HR	Historical Repository (in a GOLEM context)
HTM	Hierarchical Temporal Memory
IA	(Allen) Interval Algebra (an algebra of temporal intervals)
IRC	Imitation/Reinforcement/Correction (Learning)
LIFES	Learning-Integrated Feature Selection
LTi	Long Term Importance
MA	Mind Agent
MOSES	Meta-Optimizing Semantic Evolutionary Search
MSH	Mirror System Hypothesis
NARS	Non-Axiomatic Reasoning System
NLGen	A specific software component within OpenCog, which provides one way of dealing with Natural Language Generation
OCP	OpenCogPrime
OP	Operating Program (in a GOLEM context)
PEPL	Probabilistic Evolutionary Procedure Learning (e.g. MOSES)
PLN	Probabilistic Logic Networks
RCC	Region Connection Calculus
RelEx	A specific software component within OpenCog, which provides one way of dealing with natural language Relationship Extraction
SAT	Boolean SATisfaction, as a mathematical/computational problem
SMEPH	Self-Modifying Evolving Probabilistic Hypergraph
SRAM	Simple Realistic Agents Model
STi	Short Term Importance
STV	Simple Truth Value
TV	Truth Value
VLTI	Very Long Term Importances
WSPS	Whole-Sentence Purely-Syntactic Parsing

Chapter 1

Introduction

1.1 AI Returns to Its Roots

Our goal in this book is straightforward, albeit ambitious: to present a conceptual and technical design for a thinking machine, a software program capable of the same qualitative sort of general intelligence as human beings. It's not certain exactly how far the design outlined here will be able to take us, but it seems plausible that once fully implemented, tuned and tested, it will be able to achieve general intelligence at the human level and in some respects beyond.

Our ultimate aim is Artificial General Intelligence construed in the broadest sense, including artificial creativity and artificial genius. We feel it is important to emphasize the extremely broad potential of Artificial General Intelligence systems. The human brain is not built to be modified, except via the slow process of evolution. Engineered AGI systems, built according to designs like the one outlined here, will be much more susceptible to rapid improvement from their initial state. It seems reasonable to us to expect that, relatively shortly after achieving the first roughly human-level AGI system, AGI systems with various sorts of beyond-human-level capabilities will be achieved.

Though these long-term goals are core to our motivations, we will spend much of our time here explaining how we think we can make AGI systems do relatively simple things, like the things human children do in preschool. Chapter 3 describes a thought-experiment involving a robot playing with blocks, responding to the request "Build me something I haven't seen before". We believe that preschool creativity contains the seeds of, and the core structures and dynamics underlying, adult human level genius... and new, as yet unforeseen forms of artificial innovation.

Much of the book focuses on a specific AGI architecture, which we call CogPrime, and which is currently in the midst of implementation using the OpenCog software framework. CogPrime is large and complex and embodies a host of specific decisions regarding the various aspects of intelligence. We don't view CogPrime as the unique path to advanced AGI, nor as the ultimate end-all of AGI research. We feel confident there are multiple possible paths to advanced AGI, and that in following any of

these paths, multiple theoretical and practical lessons will be learned, leading to modifications of the ideas possessed while along the early stages of the path. But our goal here is to articulate **one** path that we believe makes sense to follow, one overall design that we believe can work.

1.2 AGI Versus Narrow AI

An outsider to the AI field might think this sort of book commonplace in the research literature, but insiders know that's far from the truth. The field of Artificial Intelligence (AI) was founded in the mid 1950s with the aim of constructing “thinking machines”—that is, computer systems with human-like general intelligence, including humanoid robots that not only look but act and think with intelligence equal to and ultimately greater than human beings. But in the intervening years, the field has drifted far from its ambitious roots, and this book represents part of a movement aimed at restoring the initial goals of the AI field, but in a manner powered by new tools and new ideas far beyond those available half a century ago.

After the first generation of AI researchers found the task of creating human-level AGI very difficult given the technology of their time, the AI field shifted focus toward what Ray Kurzweil has called “narrow AI”—the understanding of particular specialized aspects of intelligence; and the creation of AI systems displaying intelligence regarding specific tasks in relatively narrow domains. In recent years, however, the situation has been changing. More and more researchers have recognized the necessity—and feasibility—of returning to the original goals of the field.

In the decades since the 1950s, cognitive science and neuroscience have taught us a lot about what a cognitive architecture needs to look like to support roughly human-like general intelligence. Computer hardware has advanced to the point where we can build distributed systems containing large amounts of RAM and large numbers of processors, carrying out complex tasks in real time. The AI field has spawned a host of ingenious algorithms and data structures, which have been successfully deployed for a huge variety of purposes.

Due to all this progress, increasingly, there has been a call for a transition from the current focus on highly specialized “narrow AI” problem solving systems, back to confronting the more difficult issues of “human level intelligence” and more broadly “artificial general intelligence (AGI)”. Recent years have seen a growing number of special sessions, workshops and conferences devoted specifically to AGI, including the annual BICA (Biologically Inspired Cognitive Architectures) AAAI Symposium, and the international AGI conference series (one in 2006, and annual since 2008). And, even more exciting, as reviewed in Chap. 6, there are a number of contemporary projects focused directly and explicitly on AGI (sometimes under the name “AGI”, sometimes using related terms such as “Human Level Intelligence”).

In spite of all this progress, however, we feel that no one has yet clearly articulated a detailed, systematic design for an AGI, with potential to yield general intelligence at the human level and ultimately beyond. In this spirit, our main goal in this lengthy

two-part book is to outline a novel *design for a thinking machine*—an AGI design which we believe has the capability to produce software systems with intelligence at the human adult level and ultimately beyond. Many of the technical details of this design have been previously presented online in a wikibook [Goewiki]; and the basic ideas of the design have been presented briefly in a series of conference papers [GPSL03, GPPG06, Goe09c]. But the overall design has not been presented in a coherent and systematic way before this book. In order to frame this design properly, we also present a considerable number of broader theoretical and conceptual ideas here, some more and some less technical in nature.

1.3 CogPrime

The AGI design presented here has not previously been granted a name independently of its particular software implementations, but for the purposes of this book it needs one, so we’ve christened it **CogPrime**. This fits with the name “OpenCogPrime” that has already been used to describe the software implementation of CogPrime within the open-source OpenCog AGI software framework. The OpenCogPrime software, right now, implements only a small fraction of the CogPrime design as described here. However, OpenCog was designed specifically to enable efficient, scalable implementation of the full CogPrime design (as well as to serve as a more general framework for AGI R&D); and work currently proceeds in this direction, though there is a lot of work still to be done and many challenges remain.¹

The CogPrime design is more comprehensive and thorough than anything that has been presented in the literature previously, including the work of others reviewed in Chap 6. It covers all the key aspects of human intelligence, and explains how they interoperate and how they can be implemented in digital computer software. Volume 5 of this work outlines CogPrime at a high level, and makes a number of more general points about artificial general intelligence and the path thereto; then Part 2 digs deeply into the technical particulars of CogPrime. Even Part 2, however, doesn’t explain all the details of CogPrime that have been worked out so far, and it definitely doesn’t explain all the implementation details that have gone into designing and building OpenCogPrime. Creating a thinking machine is a large task, and even the intermediate level of detail takes up a lot of pages.

¹ This brings up a terminological note: At several places in this Volume and the next we will refer to the current CogPrime or OpenCog implementation; in all cases this refers to OpenCog as of late 2013. We realize the risk of mentioning the state of our software system at time of writing: for future readers this may give the wrong impression, because if our project goes well, more and more of CogPrime will get implemented and tested as time goes on (e.g. within the OpenCog framework, under active development at time of writing). However, not mentioning the current implementation at all seems an even worse course to us, since we feel readers will be interested to know which of our ideas—at time of writing—have been honed via practice and which have not. Online resources such as <http://www.opencog.org> may be consulted by readers curious about the current state of the main OpenCog implementation; though in future forks of the code may be created, or other systems may be built using some or all of the ideas in this book, etc.

1.4 The Secret Sauce

There is no consensus on why all the related technological and scientific progress mentioned above has not yet yielded AI software systems with human-like general intelligence (or even greater levels of brilliance!). However, we hypothesize that the core reason boils down to the following three points:

- Intelligence depends on the emergence of certain high-level structures and dynamics across a system's whole knowledge base;
- We have not discovered any one algorithm or approach capable of yielding the emergence of these structures;
- Achieving the emergence of these structures within a system formed by integrating a number of different AI algorithms and structures requires careful attention to the manner in which these algorithms and structures are integrated; and so far the integration has not been done in the correct way.

The human brain appears to be an integration of an assemblage of diverse structures and dynamics, built using common components and arranged according to a sensible cognitive architecture. However, its algorithms and structures have been honed by evolution to work closely together—they are very tightly inter-adapted, in the same way that the different organs of the body are adapted to work together. Due to their close interoperation they give rise to the overall systemic behaviors that characterize human-like general intelligence. We believe that the main missing ingredient in AI so far is **cognitive synergy**: the fitting-together of different intelligent components into an appropriate cognitive architecture, in such a way that the components richly and dynamically support and assist each other, interrelating very closely in a similar manner to the components of the brain or body and thus giving rise to appropriate emergent structures and dynamics. This leads us to one of the central hypotheses underlying the CogPrime approach to AGI: that **the cognitive synergy ensuing from integrating multiple symbolic and subsymbolic learning and memory components in an appropriate cognitive architecture and environment, can yield robust intelligence at the human level and ultimately beyond.**

The reason this sort of intimate integration has not yet been explored much is that it's difficult on multiple levels, requiring the design of an architecture and its component algorithms with a view toward the structures and dynamics that will arise in the system once it is coupled with an appropriate environment. Typically, the AI algorithms and structures corresponding to different cognitive functions have been developed based on divergent theoretical principles, by disparate communities of researchers, and have been tuned for effective performance on different tasks in different environments. Making such diverse components work together in a truly synergetic and cooperative way is a tall order, yet we believe that this—rather than some particular algorithm, structure or architectural principle—is the “secret sauce” needed to create human-level AGI based on technologies available today.

1.5 Extraordinary Proof?

There is a saying that “extraordinary claims require extraordinary proof” and by that standard, if one believes that having a design for an advanced AGI is an extraordinary claim, this book must be rated a failure. We don’t offer extraordinary proof that CogPrime, once fully implemented and educated, will be capable of human-level general intelligence and more.

It would be nice if we could offer mathematical proof that CogPrime has the potential we think it does, but at the current time mathematics is simply not up to the job. We’ll pursue this direction briefly in Chap. 8 and other chapters, where we’ll clarify exactly what kind of mathematical claim “CogPrime has the potential for human-level intelligence” turns out to be. Once this has been clarified, it will be clear that current mathematical knowledge does not yet let us evaluate, or even fully formalize, this kind of claim. Perhaps one day rigorous and detailed analyses of practical AGI designs will be feasible—and we look forward to that day—but it’s not here yet.

Also, it would of course be profoundly exciting if we could offer dramatic practical demonstrations of CogPrime’s capabilities. We do have a partial software implementation, in the OpenCogPrime system, but currently the things OpenCogPrime does are too simple to really serve as proofs of CogPrime’s power for advanced AGI. We have used some CogPrime ideas in the OpenCog framework to do things like natural language understanding and data mining, and to control virtual dogs in online virtual worlds; and this has been very useful work in multiple senses. It has taught us more about the CogPrime design; it has produced some useful software systems; and it constitutes fractional work building toward a full OpenCog based implementation of CogPrime. However, to date, the things OpenCogPrime has done are all things that could have been done in different ways without the CogPrime architecture (though perhaps not as elegantly nor with as much room for interesting expansion).

The bottom line is that building an AGI is a big job. Software companies like Microsoft spend dozens to hundreds of man-years building software products like word processors and operating systems, so it should be no surprise that creating a digital intelligence is also a relatively large-scale software engineering project. As time advances and software tools improve, the number of man-hours required to develop advanced AGI gradually decreases—but right now, as we write these words, it’s still a rather big job. In the OpenCogPrime project we are making a serious attempt to create a CogPrime based AGI using an open-source development methodology, with the open-source Linux operating system as one of our inspirations. But the open-source methodology doesn’t work magic either, and it remains a large project, currently at an early stage. I emphasize this point so that readers lacking software engineering expertise don’t take the currently fairly limited capabilities of OpenCogPrime as somehow a damning indictment of the potential of the CogPrime design. The design is one thing, the implementation another—and the OpenCogPrime implementation currently encompasses perhaps one third to one half of the key ideas in this book.

So we don't have extraordinary proof to offer. What we aim to offer instead are clearly-constructed conceptual and technical arguments as to why we think the CogPrime design has dramatic AGI potential.

It is also possible to push back a bit on the common intuition that having a design for human-level AGI is such an “extraordinary claim”. It may be extraordinary relative to contemporary science and culture, but we have a strong feeling that the AGI problem is not difficult in the same ways that most people (including most AI researchers) think it is. We suspect that in hindsight, after human-level AGI has been achieved, people will look back in shock that it took humanity so long to come up with a workable AGI design. As you'll understand once you've finished Part 1 of the book, we don't think general intelligence is nearly as “extraordinary” and mysterious as it's commonly made out to be. Yes, building a thinking machine is hard—but humanity has done a lot of other hard things before. It may seem difficult to believe that human-level general intelligence could be achieved by something as simple as a collection of algorithms linked together in an appropriate way and used to control an agent. But we suggest that, once the first powerful AGI systems are produced, it will become apparent that engineering human-level minds is not so profoundly different from engineering other complex systems.

All in all, we'll consider the book successful if a significant percentage of open-minded, appropriately-educated readers come away from it scratching their chins and pondering: *“Hmm. You know, that just might work”*. and a small percentage come away thinking *“Now that's an initiative I'd really like to help with!”*.

1.6 Potential Approaches to AGI

In principle, there is a large number of approaches one might take to building an AGI, starting from the knowledge, software and machinery now available. This is not the place to review them in detail, but a brief list seems apropos, including commentary on why these are not the approaches we have chosen for our own research. Our intent here is not to insult or dismiss these other potential approaches, but merely to indicate why, as researchers with limited time and resources, we have made a different choice regarding where to focus our own energies.

1.6.1 Build AGI from Narrow AI

Most of the AI programs around today are “narrow AI” programs—they carry out one particular kind of task intelligently. One could try to make an advanced AGI by combining a bunch of enhanced narrow AI programs inside some kind of overall framework.

However, we're rather skeptical of this approach because none of these narrow AI programs have the ability to generalize across domains—and we don't see how combining them or extending them is going to cause this to magically emerge.

1.6.2 Enhancing Chatbots

One could seek to make an advanced AGI by taking a chatbot, and trying to improve its code to make it actually understand what it's talking about. We have some direct experience with this route, as in 2010 our AI consulting firm was contracted to improve Ray Kurzweil's online chatbot "Ramona". Our new Ramona understands a lot more than the previous Ramona version or a typical chatbot, due to using Wikipedia and other online resources, but still it's far from an AGI.

A more ambitious attempt in this direction was Jason Hutchens' a-i.com project, which sought to create a human child level AGI via development and teaching of a statistical learning based chatbot (rather than the typical rule-based kind). The difficulty with this approach, however, is that the architecture of a chatbot is fundamentally different from the architecture of a generally intelligent mind. Much of what's important about the human mind is not directly observable in conversations, so if you start from conversation and try to work toward an AGI architecture from there, you're likely to miss many critical aspects.

1.6.3 Emulating the Brain

One can approach AGI by trying to figure out how the brain works, using brain imaging and other tools from neuroscience, and then emulating the brain in hardware or software.

One rather substantial problem with this approach is that we don't really understand how the brain works yet, because our software for measuring the brain is still relatively crude. There is no brain scanning method that combines high spatial and temporal accuracy, and none is likely to come about for a decade or two. So to do brain-emulation AGI seriously, one needs to wait a while until brain scanning technology improves.

Current AI methods like neural nets that are loosely based on the brain, are really not brain-like enough to make a serious claim at emulating the brain's approach to general intelligence. We don't yet have any real understanding of how the brain represents abstract knowledge, for example, or how it does reasoning (though the authors, like many others, have made some speculations in this regard [GMIH08]).

Another problem with this approach is that once you're done, what you get is something with a very humanlike mind, and we already have enough of those! However, this is perhaps not such a serious objection, because a digital-computer-based version of a human mind could be studied much more thoroughly than a

biology-based human mind. We could observe its dynamics in real-time in perfect precision, and could then learn things that would allow us to build other sorts of digital minds.

1.6.4 Evolve an AGI

Another approach is to try to run an evolutionary process inside the computer, and wait for advanced AGI to evolve.

One problem with this is that we don't know how evolution works all that well. There's a field of artificial life, but so far its results have been fairly disappointing. It's not yet clear how much one can vary on the chemical structures that underly real biology, and still get powerful evolution like we see in real biology. If we need good artificial chemistry to get good artificial biology, then do we need good artificial physics to get good artificial chemistry?

Another problem with this approach, of course, is that it might take a really long time. Evolution took billions of years on Earth, using a massive amount of computational power. To make the evolutionary approach to AGI effective, one would need some radical innovations to the evolutionary process (such as, perhaps, using probabilistic methods like BOA [Pel05] or MOSES [Loo06] in place of traditional evolution).

1.6.5 Derive an AGI Design Mathematically

One can try to use the mathematical theory of intelligence to figure out how to make advanced AGI.

This interests us greatly, but there's a huge gap between the rigorous math of intelligence as it exists today and anything of practical value. As we'll discuss in Chap. 8, most of the rigorous math of intelligence right now is about how to make AI on computers with dramatically unrealistic amounts of memory or processing power. When one tries to create a theoretical understanding of real-world general intelligence, one arrives at quite different sorts of considerations, as we will roughly outline in Chap. 11. Ideally we would like to be able to study the CogPrime design using a rigorous mathematical theory of real-world general intelligence, but at the moment that's not realistic. The best we can do is to conceptually analyze CogPrime and its various components in terms of relevant mathematical and theoretical ideas; and perform analysis of CogPrime's individual structures and components at varying levels of rigor.